On stereo confidence measures for global methods: evaluation, new model and integration into occupancy grids

Martim Brandão, *Member, IEEE,* Ricardo Ferreira, *Member, IEEE,* Kenji Hashimoto, *Member, IEEE,* and Atsuo Takanishi, *Member, IEEE,* and José Santos-Victor, *Member, IEEE*

Abstract—Stereo confidence measures are important functions for global reconstruction methods and some applications of stereo. In this article we evaluate and compare several models of confidence which are defined at the whole disparity range. We propose a new stereo confidence measure to which we call the Histogram Sensor Model (HSM), and show how it is one of the best performing functions overall. We also introduce, for parametric models, a systematic method for estimating their parameters which is shown to lead to better performance when compared to parameters as computed in previous literature. All models were evaluated when applied to two different cost functions at different window sizes and model parameters. Contrary to previous stereo confidence measure benchmark literature, we evaluate the models with criteria important not only to winner-take-all stereo, but also to global applications. To this end, we evaluate the models on a real-world application using a recent formulation of 3D reconstruction through occupancy grids which integrates stereo confidence at all disparities. We obtain and discuss our results on both indoors' and outdoors' publicly available datasets.

Index Terms—Stereo vision, stereo matching, confidence, uncertainty, 3D reconstruction, occupancy grids

1 Introduction

Modeling stereo matching's uncertainty is of high interest to stereo vision applications. How much confidence is to be given to a certain stereo match should be established by the right functions so that global [1], [2], [3], fusion [4], [5], [6] and progressive methods [7] are reliable. Traditionally, pixel matching costs have been used for this purpose, but it has been shown that these do not model uncertainty correctly [8]. Confidence measures of stereo are functions of stereo cost that attempt to better model match uncertainty and consequently increase performance of stereo methods. Some comparisons have been published on stereo confidence measures [8], [9] for use with winner-take-all (WTA) strategies, where only the highest-confidence estimates are considered and evaluated. However, evaluation of functions providing a confidence measure to each disparity of the disparity range is of high interest to global methods and certain global 3D reconstruction frameworks which fuse stereo information over time [4], [6]. Furthermore, performance of these functions will change depending on the choice of parameters and care should be taken to correctly estimate these before evaluation. Evaluation and proposal of confidence measures and their parameters, in terms of impact to performance of global methods, will be the focus of this article. Evaluation will be made not only on a WTA stereo paradigm, but also on the recently proposed "Cost-Curve Occupancy Grid" method [6] which fuses stereo measurements over time using the whole disparity range.

The contributions of this article are 1) A comparison of a set of models that provide a confidence measure for stereo at the whole disparity range in indoors and outdoors datasets, and an analysis of the influence of model parameters when they exist; 2) An automatic method to compute model parameters from a stereo pair without ground-truth data, based on maximum likelihood; 3) A new model, the Histogram Sensor Model (HSM), which we show to be one of the best performing; 4) A comparison of the confidence models on a real-world application - mapping of an outdoors scenario for autonomous driving. For this purpose we use an existing global occupancy grid method that integrates confidence measures at all disparities along time. Relation between results of contribution 1 and occupancy grid performance is discussed.

The structure of the article is as follows. We introduce, under a common notation, three existing and

- *M. Brandão is with the Graduate School of Advanced Science and Engineering, Waseda University, 41-304, 17 Kikui-cho, Shinjuku-ku, Tokyo 162-0044, JAPAN.*
  E-mail: contact@takanishi.mech.waseda.ac.jp
- *K. Hashimoto is with the Faculty of Science and Engineering, Waseda University.*
- *A. Takanishi is with the Department of Modern Mechanical Engineering, Waseda University; and the director of the Humanoid Robotics Institute (HRI), Waseda University.*
- *R. Ferreira and J. Santos-Victor are with the Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal.*

one new stereo confidence measures in Section 2. We then propose a method for parameter estimation of the parametric models in Section 3. We go on to briefly introduce the occupancy grid method (Section 4) and analyze the performance of the models and parameter choices in Sections 5 and 6. Conclusions are summarized in Section 7.

1.1 Background

Traditionally, uncertainty of stereo matches has been modeled by cost-functions of pixel neighborhoods, or windows. The cost function computes the cost of matching a pair of pixels between images and assumptions regard to noise distributions, continuity and local smoothness. Common cost functions include Sum of Squared Differences (SSD), Sum of Absolute Differences (SAD) and different variants of Correlation. Other more elaborate cost functions have been proposed, some of which can be implemented as a filter to the images followed by one of the previously mentioned costs [10]. For a thorough comparison of cost functions refer to [10].

Based on these cost functions several models of stereo uncertainty, or confidence measures, have been proposed since the late 1980s. Some of them assume a winner-take-all approach, refining a disparity estimate around the least cost disparity, others take all costs into consideration. Models targeting WTA stereo are usually only defined at the highest-confidence (i.e. lowest-cost) match and do not provide confidence measures on the rest of the disparity range. Examples include left-right consistency checks, uniqueness or curvature tests (how much the highest-confidence is higher than others), texture thresholds, among others. Some of these WTA confidence measures were recently reviewed in [8], [9]. Other confidence measures include statistical models that compute a variance of the disparity estimate. Some models do so by polynomial fitting [11], others by modeling disparity and texture fluctuation inside windows [12], or even by directly computing the variance of WTA disparity between different window sizes [13].

Global methods, however, usually require a likelihood function over disparity to be propagated in order to obtain a final 3D reconstruction. This asks for confidence measures that are defined along the whole disparity range and that model the confidence on each stereo match hypothesis in a reliable way. Specifically, it is not only important that the highest-confidence disparity is of high accuracy but also that when this estimate is wrong, a high confidence is still attributed to the true disparity. Figure 1 shows an example of a good confidence function, or confidence measure, in these terms. A few stereo confidence measures have been proposed that are defined at all disparities within the disparity range, although they are only evaluated at WTA disparity in recent benchmarks [8].

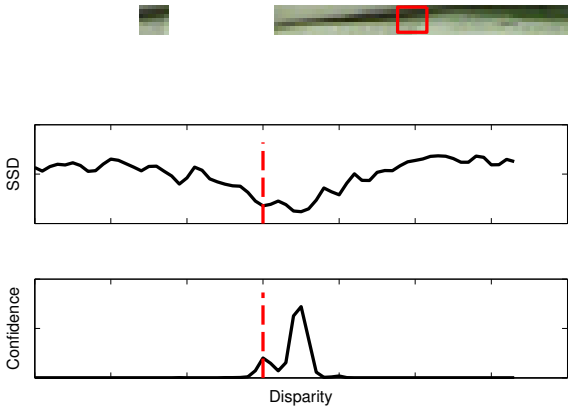

Fig. 1. Top: Matching a pixel in one image to pixels at different disparities in another image. Middle: Cost for each disparity. Bottom: Confidence measure computed from the cost values. Dashed line indicates true disparity. Even if the minimum cost is wrong, true disparity should still be attributed some confidence.

For example, in [14], Matthies and Okutomi assume normally distributed image noise and model the probability of the measured pixel differences inside a window according to that model. Sun et. al use a pixel-wise likelihood function [1] in a global stereo method, propagating these likelihoods to neighboring pixels in a Markov Random Field formulation of stereo. The cost function used was the pixel dissimilarity function proposed by Birchfield and Tomasi in [15], chosen for its invariance to image sampling. Also, Mordohai recently proposed the SAMM measure [16] which computes a confidence for each disparity based on the correlation between the left-right stereo cost curve and the self-matching (i.e. left-left) cost curve. No explicit probability distribution assumptions are made. Although promising, the function scores poorly for large support windows when used with SAD costs [16]. Merrell et. al [5] assumes costs to be normally distributed with mean equal to the best cost value and is also evaluated in [8].

Researchers have recently benchmarked several of these stereo confidence measures [8], [9], [17], [18]. Such benchmarks typically compare different methods for detection of correspondence errors [9], [17]; or evaluate whether stereo confidence measures can accurately rank matches on a WTA scenario [8], [9]. The latter make use of Receiver Operating Characteristic (ROC) curves for the evaluation, which have been frequently used in the stereo community [16], [19]. ROC curves are obtained by plotting the error-rate of a WTA strategy from the highest confidence matches, for different confidence thresholds. Using ROCs as the comparison criterion, a notable contribution to the state of the art of stereo confidence measures was made by Hu et. al [8]. In that article the authors analyze 17 different confidence functions both in

terms of detection of correct WTA matches, occlusions and performance on discontinuities. Nevertheless, the influence of parameter choice on the performance of parametric functions was not discussed. We studied this problem and present our results in this article as well, concluding that indeed parameter choice drastically influences performance both in WTA stereo and global methods. Finally, these recent benchmarks were conducted mostly for confidence measures defined only at WTA disparity. Even when measures were well defined across the whole disparity range, evaluation was only made on WTA disparity. Such evaluations are hence useful for WTA methods but less so for global methods which integrate the information at all disparities, such as those targeted in this article. They leave out possible global and semi-global stereo approaches using multiple disparity hypotheses [1], [2], [3], [6], [19], [20].

Although WTA approaches to stereo are frequently preferred due to their higher computational speed, they are more susceptible to problems with occlusions, discontinuities, noise and lack of texture. Such problems can be avoided by discarding matches that could have happened by chance (a contrario models [21]), or that are ambiguous given the confidence measure (e.g. confidently stable matching [22], training of confidence thresholds from ground-truth [23]). However, these methods come at the cost of lower density. Global methods, by considering the whole disparity range and certain geometry assumptions, have the potential to better overcome such problems. Popular examples of these methods include dynamic programming [19], optimization methods using Markov network representations of stereo matching [1], [2], [3], among others.

Furthermore, we recently showed that occupancy grid algorithms using stereo sensors can also improve performance by integrating confidence measures at all disparities instead of WTA disparity alone [24], [6]. This integration of several stereo pairs into a final occupancy grid was the chosen application in the present article for confidence measure evaluation. Such is a typical scenario found in real-world robotics applications and autonomous driving applications, which are usually approached using grid-based methods [4], [6], [25], [23]. Inclusively, recent work has provided the community with urban driving datasets including stereo and laser rangefinder data which can be used as ground-truth [26]. The existence of such datasets also asks for an evaluation of stereo confidence functions and their global integration in time in such challenging scenarios.

2 Stereo Confidence Measures

We consider two images $I_1(x, y)$ and $I_2(x, y)$ coming from the same underlying image $I(x, y)$, displaced along the x axis with added Gaussian noise. Therefore,

$$I_2(x, y) - I_1(x + d(x, y), y) = \mathcal{N}(0, \sigma_i^2) \tag{1}$$

where $\mathcal{N}(0, \sigma_i^2)$ represents Gaussian white noise with variance equal to the sum of noise variances of each image $\sigma_i^2 = \sigma_1^2 + \sigma_2^2$. Here $d(x, y) \in \{0, 1, ..., D-1\}$ represents the disparity at each pixel. We define also a window with $M \times N$ pixels where (x, y) is the anchor pixel in the center of the window.

Different confidence measures model stereo matches differently. For example, one can model the probability of a disparity value $d(x, y)$ conditioned on a cost function of the pixels inside a window, but another option is to condition disparity on the whole set of pixel differences inside that window. We then define for each pixel (x, y) a matrix of measurements $E \in \mathbb{R}^{S \times D}$, where the D columns are disparity hypotheses and the rows are measurements used for the stereo confidence model (e.g. $S = 1$ for a single cost value per disparity, or $S = MN$ pixel differences per disparity). We will use the notation $E_{:,d}$ to represent all rows taken at disparity d. We will also refer to the disparity with minimum cost by $d_{mincost}$. Finally, in this work we assume independence of measurements at different disparities such that

$$p(E) = \prod_d p(E_{:,d}). \tag{2}$$

In this article we will deal with a special class of stereo confidence measures defined along the whole disparity range such that

$$C(d) = \frac{p(E_{:,d} \mid d)}{\sum_{d'} p(E_{:,d'} \mid d')} \tag{3}$$

is the confidence of assigning disparity d to a certain pixel, and $p(E_{:,d} \mid d)$ is the probability density of measurements assuming d is the true disparity. Such formulation is used implicitly in other benchmarks [8] and will also be convenient for the integration into probabilistic frameworks described in Section 4.

We will evaluate and compare different confidence measures with two different stereo cost functions:

- Sum of Squared Differences (SSD)
- Sum of Absolute Differences (SAD) using Birchfield and Tomasi's pixel dissimilarity function [15], which we will call BTSAD.

These are widely used cost functions, adopted by recent computer vision libraries [27] for local and global stereo methods. The implementations used in this work were those found in OpenCV [27], which also apply a 9x9 Sobel filter as a prefilter to the images. Sobel prefiltering is a common procedure seen in other stereo methods as well (e.g. [28]).

2.1 Matthies' model

Matthies and Okutomi [14] propose a probabilistic model of stereo that assumes pixel differences inside

a window to be i.i.d. and zero-mean Gaussian distributed. The joint probability of all pixel differences is given by

$$p(E_{:,d} \,|\, d) \overset{i.i.d.}{=} \prod_s p(E_{s,d} \,|\, d) \propto exp\left(-\frac{1}{2\sigma_{Mat}^2} \sum_s E_{s,d}^2\right), \quad (4)$$

where $E \in \mathbb{R}^{S \times D}$ with $S = MN$. Each element $E_{s,d}$ holds one of the MN pixel differences inside a window at disparity d. Note that the joint distribution is related to a SSD ($\sum_s E_{s,d}^2$). Similarly to recent literature [8], we normalize the SSD by the number of window pixels[1] by setting $\sigma_{Mat}^2 = MN\sigma_i^2$.

To obtain a similar model for a SAD cost function we can assume the i.i.d. pixel differences to follow a zero-mean Laplace distribution. The joint distribution is then given by

$$p(E_{:,d} \,|\, d) \overset{i.i.d.}{=} \prod_s p(E_{s,d} \,|\, d) \propto exp\left(-\frac{1}{b_{Mat}} \sum_s |E_{s,d}|\right). \quad (5)$$

In this case the joint distribution is related to a SAD ($\sum_s |E_{s,d}|$). Likewise the SSD case and since it lead us to better performance, we set $b_{Mat} = MNb_i$ where b_i is the parameter of the zero-mean Laplacian of single pixel differences.

2.2 Merrell's model

Merrel et. al [5] assume costs themselves to be normally distributed. The mean is set to the minimum cost of the corresponding pixel and variance is a parameter σ_{Mer}^2. Confidence is in this case defined by

$$p(E_{1,d} \,|\, d) \propto exp\left(-\frac{(E_{1,d} - E_{1,d\_mincost})^2}{2\sigma_{Mer}^2}\right), \quad (6)$$

where $E \in \mathbb{R}^{1 \times D}$ and each element $E_{1,d}$ is a window cost value, e.g. $E_{1,d}$ = SSD or BTSAD.

2.3 The exponential distribution

The exponential model [1], [2], [3] assumes costs to be exponentially distributed and is given by

$$p(E_{1,d} \,|\, d) \propto exp\left(-\frac{E_{1,d}}{\mu}\right), \quad (7)$$

where $E \in \mathbb{R}^{1 \times D}$ and each element $E_{1,d}$ is a window cost value, e.g. $E_{1,d}$ = SSD or BTSAD. Note that this model's expression is similar to Matthies'. However, while the exponential model is a pdf of the cost values, Matthies' is a joint pdf of all window pixel differences.

Note also that in other literature μ is often omitted from the equations, thus $\mu = 1$ is often assumed. The underlying problem of that assumption is that, for $\mu << E_{1,d}$ equation (7) will approximate $min(E_{1,d})$ and thus $p(E_{1,d\_mincost} \,|\, d_{mincost}) = 1$ will hold for all $d_{mincost}$. Such choice of parameter could hence lead to low performance of the confidence measure.

1. Note that the original model [14] sets $\sigma_{Mat}^2 = \sigma_i^2$. While the normalization by MN was not used in that publication, we still refer to the model as used in this article as "Matthies' model" for acknowledgment.

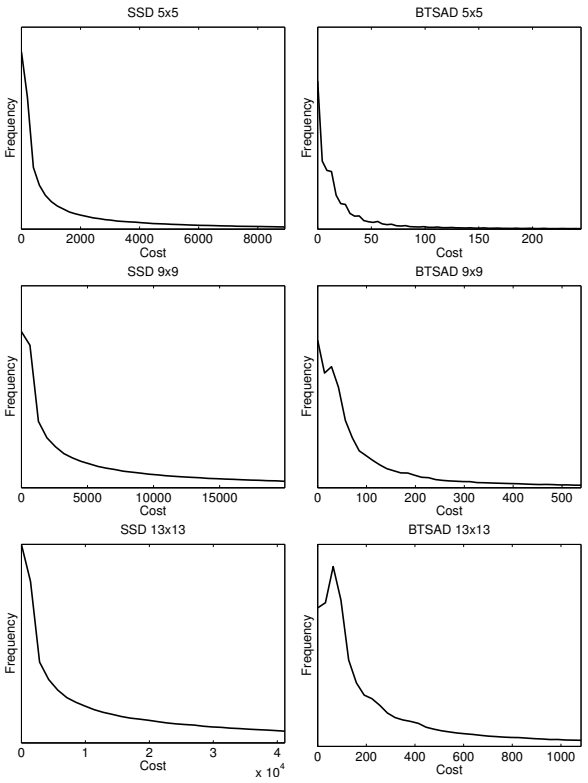

Fig. 2. Distribution of costs at true disparity (E_{1,d^*}) for SSD (left) and BTSAD (right) cost functions on a 5x5, 9x9 and 13x13 window. Horizontal axis represents the values of E_{1,d^*}.

2.4 New confidence measure: Histogram Sensor Model (HSM)

We finally propose our new confidence measure - the HSM - which consists of a histogram trained with costs at true disparity. Confidence is modeled from the cost values and as such $E \in \mathbb{R}^{1 \times D}$. In Figure 2, we show these histograms for SSD and BTSAD costs with different window sizes, taken from true disparity d of all images in the 2003 and 2006 Middlebury datasets. We populated the histograms with costs measured at all un-occluded pixels of all images, while true disparity was retrieved from the ground-truth disparity maps provided by the datasets. The dimension of bins was chosen at $3.5\sigma_h/N^{1/3}$ according to Scott's normal reference rule [29], where σ_h represents the standard deviation of the costs and N the number of samples.

Stereo confidence is in this case defined as

$$p(E_{1,d} \,|\, d) \propto hist(E_{1,d}), \quad (8)$$

where $E_{1,d}$ is a window cost value, e.g. $E_{1,d}$ = SSD or BTSAD, and $hist(E_{1,d})$ refers to the frequency of the

histogram bin associated with $E_{1,d}$.

3 Parameter Estimation

The parametric confidence measures introduced so far depend on the estimation of a probability distribution's parameter (σ^2_{Mat}, σ^2_{Mer}, μ). In this section we propose to estimate the parameters in a systematic way without ground-truth data, from each stereo pair being matched: through maximum likelihood (ML) estimation of the distribution's parameters computed directly from cost values. The method does not require ground-truth data but assumes cost functions provide relatively low error-rates (low number of bad pixels). To achieve this, *in our study we compute ML parameters from costs at all image pixels where left-right disparity consistency is verified.*

In a nutshell, we: 1) Compute cost values at all pixels and disparities; 2) Compute $d_{mincost}$ and perform a left-right disparity consistency check; 3) For all (x,y) with consistent disparities we compute the mean and variance of the costs at $d_{mincost}$; 4) Compute model parameters from those means or variances.

3.1 Matthies' model

Matthies' model for the SSD cost function assumes pixel differences to be zero-mean Gaussian. The Gaussian's parameter σ_i^2 can be computed by maximum likelihood from the variance of the data. For convenience we estimate this variance from the SSD cost values instead of the individual pixel differences. We do this by the following heuristic[2], which we found best performing:

$$\hat{\sigma}_i^2 = \frac{\sqrt{Var_{x,y}(SSD(x,y,d_{mincost}(x,y)))}}{MN\sqrt{2}}. \quad (9)$$

As mentioned in Section 2.1 we set $\hat{\sigma}^2_{Mat} = MN\hat{\sigma}_i^2$, which is effectively eliminating the MN normalization in (9).

On a SAD (or BTSAD) cost function, we assume pixel differences are zero-mean Laplace-distributed, for which the maximum likelihood parameter is the mean of the absolute value of the data. As done in the SSD case, we compute this estimate from the cost values themselves:

$$\hat{b}_i = \frac{Mean_{x,y}(BTSAD(x,y,d_{mincost}(x,y)))}{MN}, \quad (10)$$

and we set $\hat{b}_{Mat} = MN\hat{b}_i$. Please note that using this normalization makes $\hat{b}_{Mat}$ equal to the costs' mean, leading to the same model expression and parameter as the exponential model (see (7) (12)). In this article, results obtained by maximum likelihood will then be the same for BTSAD Matthies' and the BTSAD exponential models.

2. Note that from the moments of the normal distribution we know that a variable X^2 has variance $2\sigma^4$ for $X = \mathcal{N}(0,\sigma^2)$. We compute the variance of an SSD by $Var(\sum_{s=1}^{MN} E_s^2) = 2\sigma_i^4 MN(1+\rho(MN-1))$, where ρ is the average correlation between the squared pixel differences E_s^2. Our heuristic assumes $\rho = 1$. While the original i.i.d. assumption of the model [14] would lead to $\rho = 0$, assuming $\rho = 1$ lead us to better performance results. Finally, note that another option for estimating σ_i^2 would be $\sigma_i^2 = Mean(\sum_{s=1}^{MN} E_s^2)/(2MN)$, which would make the estimated model's expression equal to that of the exponential.

3.2 Merrell's model

Merrell's model is a Gaussian distribution of costs with mean $E_{1,d\_mincost}$. The maximum likelihood parameter is estimated from the variance of the data,

$$\hat{\sigma}^2_{Mer} = Var_{x,y}(E_{1,d\_mincost}(x,y)), \quad (11)$$

where $E_{1,d\_mincost}$ is an SSD or BTSAD.

3.3 The exponential distribution

Given an exponential distribution of costs, the maximum likelihood estimate of the distribution's parameter μ is given by

$$\hat{\mu} = Mean_{x,y}(E_{1,d\_mincost}(x,y)), \quad (12)$$

where $E_{1,d\_mincost}$ is an SSD or BTSAD.

4 Integrating Stereo into Occupancy Grids using Confidence Measures

Consider a grid of cells which can be in one of two states: occupied O or free $\overline{O}$. The objective of an occupancy grid algorithm is to compute or update the probabilities $p(O_i|z_{0...t}, x_{0...t})$ for each cell $i \in 1,2,...,C$, at each time instant t, given measurements $z_{0...t}$ and sensor locations $x_{0...t}$ until time t. This is implemented as a Bayes filter at each cell, which updates occupancy probabilities every time a new measurement is taken [30].

In this article we use a Cost-Curve Occupancy Grid [6] to compute occupancy at each cell from stereo cost measurements at the whole disparity range. The method computes occupancy of cell i as

$$\begin{aligned} P(O_i|E) = P(O_i|V_i,E)P(V_i|E) + \\ P(O_i|\overline{V}_i,E)(1-P(V_i|E)), \end{aligned} \quad (13)$$

where the event $V_i = \overline{O}_{i-1},...,\overline{O}_2,\overline{O}_1$ represents visibility of cell i. For the sake of readability and compactness, the equations shown here are for a one-dimensional grid aligned with the sensor - correspondent to the intersection of a camera ray with the three-dimensional grid. Also, the order of cells is reversed from that of pixel disparity: for example $i = 1$ is the closest cell to the camera, equivalent to $d = D - i = D - 1$.

In the original paper [6], which the interested reader should refer to, it is demonstrated that

$$P(V_i|E) = \prod_{j=1...i-1} P(\overline{O}_j|V_j,E), \quad (14)$$

$$P(O_i|V_i,E) = \frac{p(E|O_i,V_i)P(O_i,V_i)}{P(V_i|E)p(E)}, \quad (15)$$

$$P(V_i|E)p(E) = \sum_{j=i...C} p(E|O_j, V_j)P(O_j, V_j), \quad (16)$$

$$P(O_i|V_i, E) = \frac{p(E_{:,D-i}|O_i, V_i)}{\sum_{j=i...C} p(E_{:,D-j}|O_j, V_j)}. \quad (17)$$

Note that (17) is similar to our definition of stereo match confidence (3) if disparity is seen as a position (i.e. cell) which is both occupied and visible.

As discussed in [6], the method makes the following assumptions:

- A target surface exists for any 1D grid, or in other words, there exists at least one occupied cell. Thus $P(V_{C+1}) = 0$ and $P(V_{C+1}|E) = 0$;
- The target is equally probable to be at any of the cells along the 1D grid. Thus $P(O_i, V_i) = 1/C \ \forall_i$;
- Measurements E can give no information about occupancy on invisible cells $\overline{V}_i$. Thus $P(O_i|\overline{V}_i, E) = P(O_i|\overline{V}_i)$, which corresponds to a prior on world geometry. In our work we model this prior as a constant 0.5 for all i, so that occupied and free cells are equally probable. Thus $P(O_i|\overline{V}_i) = 0.5 \ \forall_i$;
- Measurements are independent between disparities (see (2)).
- $p(E_{:,d})$ is uniform.
- Occupancy or visibility on a cell i gives no information on match measurements taken on other cells. Thus $p(E_{:,D-k}|O_i, V_i) = p(E_{:,D-k}) \ \forall_{k \neq i}$;

5 Experimental Results in Stereo

In this section we make use of stereo datasets and their ground-truth data to evaluate and compare the introduced stereo confidence measures. We base our comparison on two criteria:

1. Performance on a WTA strategy (selecting maximum confidence disparity at each pixel). For easy comparison with other literature, we make use of ROC curves [19], [16], [8]. These curves are obtained by plotting the error-rate of a WTA strategy from the highest confidence matches, for different confidence thresholds. The area under this curve, AUC, is used to measure the quality of the function as a confidence measure. Concretely, whether correct matches are given higher confidence than incorrect ones. *Lower values of AUC mean better performance.*

2. We consider the cases where WTA disparity is different from true disparity by more than one pixel (we will call these "bad pixels"). We compute, at all bad pixels, the sum of the confidence attributed to a neighborhood around ground-truth disparity d^* given by the dataset: $C(d \in GT)_{badpx} = \sum_{d \in GT} C(d)$. Here GT represents the interval $[d^* - 1; d^* + 1]$. A single performance indicator for each image is then given by the average of $C(d \in GT)_{badpx}$ over all bad pixels. *Higher values of $C(d \in GT)_{badpx}$ indicate higher probability given to true disparity and, as we will argue, better performance of some global algorithms.*

We evaluated all models in two sets of data:

1. Indoors set: 23 stereo pairs (all pairs from Middlebury 2003 and 2006 [31], [32], [33])

2. Outdoors set: 10 stereo pairs (KITTI stereo dataset [26], first 10 images).

For each set, the AUC and $C(d \in GT)_{badpx}$ results are averaged from all its stereo pairs and occluded pixels are excluded. The images were used in grayscale. As cost functions we used SSD, and SAD with BT pixel differences (BTSAD) on window sizes 5x5, 9x9 and 13x13, after prefiltering the images with a Sobel 9x9 filter (OpenCV implementation [27]). This prefilter is adopted in several stereo methods (e.g. [27], [28]) and we also found both AUC and $C(d \in GT)_{badpx}$ performance to improve significantly with prefiltering for all models.

5.1 Parametric models: the influence of parameter choice

For the parametric functions introduced in Section 2, we evaluated the influence of parameter choice on the two mentioned performance criteria (i.e. AUC and $C(d \in GT)_{badpx}$). In Figure 3 we show the performance curves obtained for different window sizes, cost functions and confidence measures. Results are shown for four of the indoors stereo pairs. Other stereo pairs have similar curves, although we do not display all to keep figures understandable. The results show that performance of the confidence measures, with respect to parameter choice, has one clear maximum followed by a slow exponential decay of performance. However, a performance "cliff" exists as the parameter tends to zero (i.e. is under-estimated). One important observation is that $\mu = 1$ or $\mu = MN$, common parameter choices for the exponential model [8], could easily fall into the "performance cliff" by underestimating noise, thus drastically reducing performance. We believe this to be the reason why that model scores poorly in recent benchmarks [8] (it is there called Negative Entropy Measure). Furthermore, we argue that measuring parameter sensitivity through an analysis such as the one in Figure 3 or similar, should be used in future benchmarks and confidence measure proposals for more complete evaluations.

Another interesting observation is that these parameter performance curves have some inter-image variability. For each combination of cost function and window size, we computed the standard-deviation of the optimal parameter values across the 23 images of the indoors set. The average standard deviation of parameters was 131% when optimizing AUC and 84% when optimizing $C(d \in GT)_{badpx}$. On the other hand, optimal parameters also highly depend on the chosen cost function: for a fixed image the average standard-deviation across all combinations of cost function and window size was 352% in the AUC case and 338%

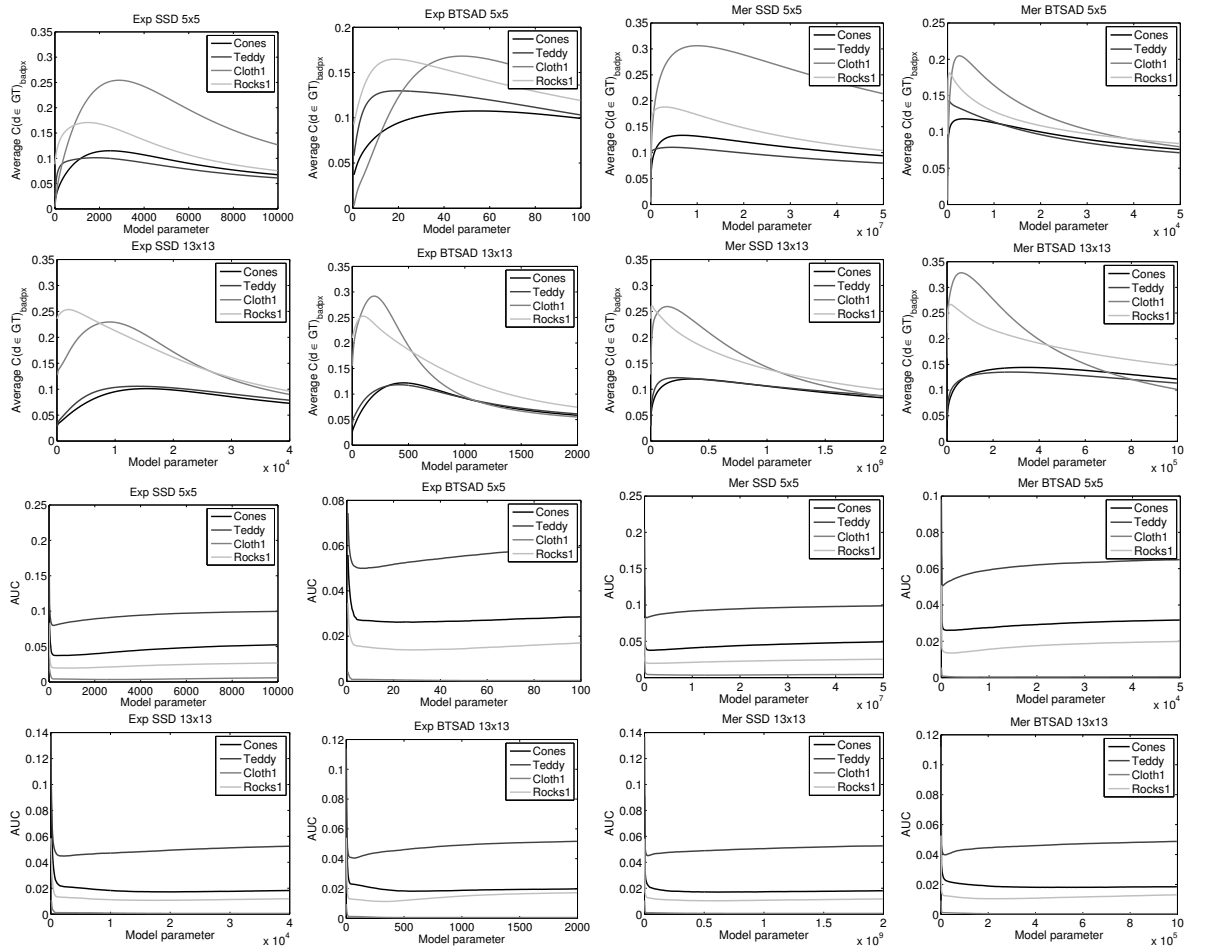

Fig. 3. The parametric models' cliff-maximum-and-tail of performance. Both $C(d \in GT)_{badpx}$ (first 2 rows) and AUC (last 2 rows) are shown for the exponential and Merrell models. Results with the different cost functions and window sizes are shown. Note how the curves and optimal parameters vary both between images and cost functions. Figures for Matthies' model are not shown since they can be obtained by linearly rescaling the horizontal axis of the exponential model's figures (see equations (4), (5) and (7)).

in the $C(d \in GT)_{badpx}$ case. Even the fact that a prefilter is applied to the images, in our case the commonly used Sobel filter [27], [28], leads to an average displacement of the parameter with optimal AUC by 60% or optimal $C(d \in GT)_{badpx}$ by 167%. Figure 4 shows such a comparison, taken from the Cones image in the indoors set. Still, note that the AUC curves are relatively flat after the performance cliff and so optimal parameter variabilty does not pose a problem as long as parameters are not strongly under or overestimated.

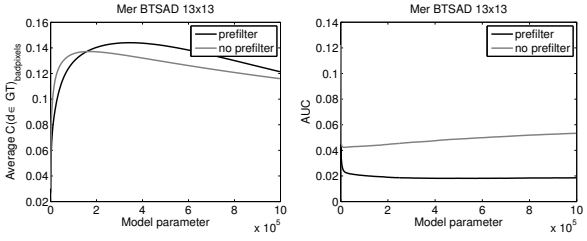

Fig. 4. Performance of models with parameter values changes with prefiltering conditions. Results obtained from the Cones image of the indoors set.

Such performance variability between image conditions and between cost function options has strong implications for researchers working on stereo. During the design stage of a stereo algorithm, such as the experimentation with different cost definitions, prefiltering options and different datasets, the optimal value of the confidence measure's parameter should be recomputed each time. In Hu et. al's important contribution to confidence measure benchmarking [8], the authors compute an optimal parameter value for each measure on a subset of the images in the dataset: which requires recomputing all confidences and a performance value (e.g. AUC) for each parameter sample during an optimization process. The parameters were there selected such that they lead on average to high performance within a subset of the dataset images,

although the procedure is not described in detail. Besides the fact that averaging solves inter-image variability sub-optimally, such methodology (of optimal parameter estimation from datasets with ground-truth) could be a bothersome process when designing a stereo algorithm and considering a large number of cost function or prefiltering options. Automatic, fast estimation of stereo confidence parameters for a given image and cost function design, for example through maximum likelihood as done in this article, is then of high importance.

5.2 Parametric models: parameter estimation

Optimal parameters for the confidence measures can only be computed when ground-truth disparity is available. Practically, on unknown stereo pairs, stereo methods have to either assume certain fixed parameter values (as discussed previously), or automatically estimate them from each image without ground-truth data. In this section we evaluate two different parameter estimation strategies for the parametric models:

- Fixed parameters, computed using a slow offline optimization procedure on training datasets where ground-truth is available. Methodology used was similar to [8]: we estimated parameters by averaging the optimal parameters across train set images. For each image in the indoors set we first computed densely sampled parameter-performance curves such as the ones shown in Figure 3, and then averaged the curves' optima across all images. We will call these "average best performing" (ABP) parameters.
- Per-stereo-pair, maximum likelihood (ML) parameter estimation as proposed in this article, which does not require any ground-truth data. We will call these "ML" parameters.

Table 1 shows the ABP parameters that we used in this article, computed from the indoors set. Since these can be chosen to optimize either AUC or $C(d \in GT)_{badpx}$, we display both in the table. As already discussed in Section 5.1, ABP parameters optimizing AUC (column "minAUC") have more variability than those optimizing $C(d \in GT)_{badpx}$ (column "maxC"). This suggests that a strategy of offline selection of parameters by averaging on a training set could be more reliable if the criterion being optimized is C.

We then computed the AUC and $C(d \in GT)_{badpx}$ metrics for each model using ML and ABP parameters. Table 2 shows the average and standard deviation of the distances between the obtained and the optimal performance taken from all 23 images of the indoors set. The table compares two situations: a typical scenario where ground-truth (GT) is not available on the image set, and another when it is available. In the "No GT" scenario, ABP parameters are computed from a different set (same images but without the use of image prefiltering with a Sobel prefilter). It is noticeable how in both situations ML parameters lead to values of AUC and $C(d \in GT)_{badpx}$ which are similar but slightly closer to the optimal value than ABP. This was expected from the analysis in Section 5.1 where we discussed high variability of optimal parameters, thus again stressing the importance of ML estimation or the use of parameter-insensitive confidence measures. The table also shows results obtained with the ML method ran on GT disparity instead of WTA (see columns ML-GT). It performed similarly to the no-ground-truth version and better than ABP on average. Importantly, these results mean that the tedious process of obtaining datasets with ground-truth for model training is unnecessary. Model parameters can be computed using our proposed ML strategy, without ground-truth data. Naturally, ABP had slightly higher performance when trained with GT than in the "No GT" condition.

TABLE 1
Average best performing parameters computed from the indoors set (total 23 images)

Cost	Model	minAUC param	maxC param
SSD 5x5	Mat	$2.95 \cdot 10^2 \pm 151\%$	$5.99 \cdot 10^2 \pm 92\%$
SSD 9x9	Mat	$1.91 \cdot 10^3 \pm 126\%$	$2.36 \cdot 10^3 \pm 47\%$
SSD 13x13	Mat	$4.17 \cdot 10^3 \pm 117\%$	$4.83 \cdot 10^3 \pm 42\%$
SSD 5x5	Mer	$2.59 \cdot 10^6 \pm 197\%$	$3.49 \cdot 10^6 \pm 103\%$
SSD 9x9	Mer	$5.49 \cdot 10^7 \pm 146\%$	$3.92 \cdot 10^7 \pm 65\%$
SSD 13x13	Mer	$2.82 \cdot 10^8 \pm 147\%$	$1.55 \cdot 10^8 \pm 59\%$
SSD 5x5	Exp	$5.94 \cdot 10^2 \pm 150\%$	$1.20 \cdot 10^3 \pm 93\%$
SSD 9x9	Exp	$3.67 \cdot 10^3 \pm 130\%$	$3.15 \cdot 10^3 \pm 98\%$
SSD 13x13	Exp	$8.27 \cdot 10^3 \pm 118\%$	$8.70 \cdot 10^3 \pm 56\%$
BTSAD 5x5	Mat	$1.18 \cdot 10^1 \pm 106\%$	$1.18 \cdot 10^1 \pm 88\%$
BTSAD 9x9	Mat	$5.64 \cdot 10^1 \pm 110\%$	$4.24 \cdot 10^1 \pm 94\%$
BTSAD 13x13	Mat	$1.12 \cdot 10^2 \pm 105\%$	$1.40 \cdot 10^2 \pm 67\%$
BTSAD 5x5	Mer	$1.88 \cdot 10^3 \pm 173\%$	$1.25 \cdot 10^3 \pm 126\%$
BTSAD 9x9	Mer	$3.89 \cdot 10^4 \pm 130\%$	$1.94 \cdot 10^4 \pm 124\%$
BTSAD 13x13	Mer	$1.81 \cdot 10^5 \pm 132\%$	$1.91 \cdot 10^5 \pm 101\%$
BTSAD 5x5	Exp	$2.37 \cdot 10^1 \pm 106\%$	$2.37 \cdot 10^1 \pm 88\%$
BTSAD 9x9	Exp	$1.13 \cdot 10^2 \pm 110\%$	$8.49 \cdot 10^1 \pm 94\%$
BTSAD 13x13	Exp	$2.24 \cdot 10^2 \pm 105\%$	$2.81 \cdot 10^2 \pm 67\%$

To exemplify the better results of ML seen in Table 2, we also compare the shape of $C(d)$ at a given pixel of Middlebury's Teddy image which favors the ML method. In this example, shown in Figure 5, Merrell's model with ABP parameters behaves in a unimodal way (i.e. single maximum), which exemplifies the effect of the "performance-cliff". We remind that as σ tends to 0, a normalized $exp(-\frac{x}{\sigma})$ becomes an approximation to $min(x)$, thus leading to a confidence of 1 on the best match and 0 otherwise. The model using ML parameters has two maxima: one on WTA disparity and another on ground-truth.

5.3 All models: evaluation of winner-take-all confidence

We evaluated each models' performance, including the HSM's, in the indoors and outdoors set using the two parameter selection strategies already discussed.

TABLE 2

On average, how close to optimal performance do models get? Distances computed as $|AUC_{Method}(img) - minAUC(img)|/minAUC(img)$ and $|C_{Method}(img) - maxC(img)|/maxC(img)$ averaged over all indoors images. ABP are average best performing parameters trained on the same image set given GT disparity; ABP-DS are average best performing parameters trained on a different set - same images different filtering conditions; ML parameters computed for each image given WTA disparity; ML-GT parameters computed using the same method on ground-truth disparity.

	Distance to minAUC				Distance to maxC			
	No GT available		GT available		No GT available		GT available	
Model	ML	ABP-DS	ML-GT	ABP	ML	ABP-DS	ML-GT	ABP
Mat SSD	**0.08 ± 0.07**	0.12 ± 0.22	**0.11 ± 0.09**	0.11 ± 0.13	**0.11 ± 0.14**	0.19 ± 0.15	0.19 ± 0.16	**0.11 ± 0.12**
Mat BTSAD	**0.10 ± 0.22**	0.14 ± 0.29	**0.08 ± 0.17**	0.11 ± 0.14	**0.11 ± 0.09**	0.14 ± 0.10	**0.09 ± 0.08**	0.11 ± 0.11
Mer SSD	**0.06 ± 0.05**	0.12 ± 0.22	**0.06 ± 0.06**	0.09 ± 0.08	**0.04 ± 0.05**	0.10 ± 0.09	**0.07 ± 0.09**	0.07 ± 0.10
Mer BTSAD	**0.13 ± 0.27**	0.15 ± 0.29	**0.09 ± 0.18**	0.11 ± 0.10	**0.10 ± 0.08**	0.13 ± 0.08	**0.09 ± 0.08**	0.14 ± 0.17
Exp SSD	**0.06 ± 0.05**	0.12 ± 0.22	**0.08 ± 0.06**	0.11 ± 0.13	**0.12 ± 0.13**	0.19 ± 0.15	0.15 ± 0.15	**0.11 ± 0.12**
Exp BTSAD	**0.10 ± 0.22**	0.14 ± 0.29	**0.08 ± 0.17**	0.11 ± 0.14	**0.11 ± 0.09**	0.14 ± 0.10	**0.09 ± 0.08**	0.11 ± 0.11

TABLE 3

Performance in AUC for all models and window cost functions, averaged over a test set. Note: lower AUC is better. ABP are average best performing parameters computed from the indoors set using ground-truth; AGT are average ground-truth histograms as proposed in Section 2.4 i.e. HSMs trained on the whole indoors set using ground-truth; ML parameters are estimated for each image from WTA disparity, without ground-truth. Optimal AUC values are shown for comparison and were computed by a slow offline optimization procedure given ground-truth (minimum AUC across all parametric models and whole parameter space).

Test set: indoors (ABP/AGT is trained on the same set and requires GT disparity)									
	Optimal AUC	Mat		Mer		Exp		HSM	
Cost	(parametric)	ABP	ML	ABP	ML	ABP	ML	AGT	ML
SSD 5x5	0.083	0.087	0.088	0.091	0.087	0.087	**0.086**	0.088	0.106
SSD 9x9	0.058	0.063	0.063	0.065	0.063	0.063	0.062	**0.062**	0.085
SSD 13x13	0.056	0.060	0.061	0.062	0.060	0.060	0.060	**0.060**	0.084
BTSAD 5x5	0.066	0.069	0.067	0.070	0.068	0.069	0.067	**0.058**	0.065
BTSAD 9x9	0.051	0.055	0.054	0.056	0.054	0.055	0.054	**0.045**	0.058
BTSAD 13x13	0.050	0.054	0.053	0.056	0.053	0.054	0.053	**0.046**	0.064
Test set: outdoors (ABP/AGT is trained on a different set - indoors)									
	Optimal AUC	Mat		Mer		Exp		HSM	
Cost	(parametric)	ABP-DS	ML	ABP-DS	ML	ABP-DS	ML	AGT-DS	ML
SSD 5x5	0.223	0.230	0.233	0.233	0.229	0.230	0.232	**0.225**	0.256
SSD 9x9	0.175	0.180	0.184	0.183	0.181	0.180	0.183	**0.176**	0.230
SSD 13x13	0.202	0.205	0.207	0.206	0.206	0.205	0.207	**0.200**	0.273
BTSAD 5x5	0.147	0.152	0.153	0.155	**0.152**	0.152	0.153	0.153	0.157
BTSAD 9x9	0.117	**0.121**	0.123	0.124	0.121	0.121	0.123	0.122	0.136
BTSAD 13x13	0.145	0.148	0.149	0.149	0.148	0.148	0.149	**0.145**	0.168

TABLE 4

Performance in $C(d \in GT)_{badpx}$. Note: higher C is better. See description in Table 3.

Test set: indoors (ABP/AGT is trained on the same set and requires GT disparity)									
	Optimal C	Mat		Mer		Exp		HSM	
Cost	(parametric)	ABP	ML	ABP	ML	ABP	ML	AGT	ML
SSD 5x5	0.108	0.083	0.090	0.097	**0.097**	0.083	0.090	0.077	0.083
SSD 9x9	0.091	0.076	0.072	0.084	**0.086**	0.076	0.074	0.061	0.066
SSD 13x13	0.101	0.086	0.073	0.093	**0.094**	0.086	0.073	0.060	0.072
BTSAD 5x5	0.109	0.087	0.086	0.088	**0.095**	0.087	0.086	0.076	0.094
BTSAD 9x9	0.099	0.084	0.083	0.090	**0.090**	0.084	0.083	0.067	0.085
BTSAD 13x13	0.112	0.095	0.094	**0.104**	0.103	0.095	0.094	0.070	0.088
Test set: outdoors (ABP/AGT is trained on a different set - indoors)									
	Optimal C	Mat		Mer		Exp		HSM	
Cost	(parametric)	ABP-DS	ML	ABP-DS	ML	ABP-DS	ML	AGT-DS	ML
SSD 5x5	0.065	0.053	0.049	0.052	**0.062**	0.053	0.050	0.031	0.043
SSD 9x9	0.059	0.047	0.036	0.045	**0.051**	0.047	0.036	0.025	0.028
SSD 13x13	0.046	0.037	0.029	0.036	**0.039**	0.037	0.029	0.022	0.020
BTSAD 5x5	0.084	0.063	0.060	0.055	**0.072**	0.063	0.060	0.040	0.061
BTSAD 9x9	0.079	0.055	0.045	0.048	**0.061**	0.055	0.045	0.030	0.050
BTSAD 13x13	0.069	0.048	0.039	0.043	**0.051**	0.048	0.039	0.027	0.040

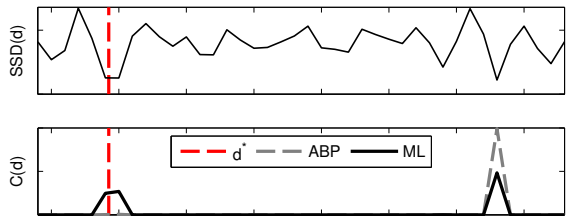

Fig. 5. $C(d)$ given Merrell's model with ABP and ML parameters. Dashed red line indicates true disparity d^* as indicated by the dataset. Results taken from pixel (364,150) of the Teddy image, as an example of ML's better performance seen in Table 2. ML does not require ground-truth and leads here to higher $C(d^*)$.

In this section we focus on the AUC criterion. We remind that AUC measures whether higher confidence WTA assignments are more likely to be correct assignments or not. The models' AUC, averaged across all images in each dataset, is shown in Table 3. Each model's performance is shown with ML and ABP parameters. In case of the HSM, we also compare two versions of the model, roughly corresponding to ML and ABP. The first version is a no-ground-truth single-stereo-pair model to which we will call "ML HSM". This histogram is trained from WTA disparity costs where left-right disparity is consistent, for each stereo pair. The second is the ground-truth-trained model as described in Section 2.4, computed from the costs at true disparity of all stereo pairs in the indoors set. We refer to it as "average ground-truth" (AGT) HSM.

Table 3 also shows the optimal AUC across parametric models, for each cost function. These values were obtained by a slow offline optimization procedure given ground-truth data, searching the minimum AUC across all parametric models and whole parameter space for each image. Values shown in the table are the average over all test set's images.

Arguably the most noticeable result is that the AGT HSM model ranks 1st in most conditions, both indoors (where it is trained) and outdoors. This indicates the HSM model to be a good choice when training on a dataset with ground-truth is acceptable. Expectedly, a histogram can better model the real distribution of costs than the parametric models here compared - we remind that distributions in Figure 2 are not purely exponential or Gaussian. This can also be seen clearly in the table results (indoors set, BTSAD cost function) where the HSM performs better than the parametric models' maximum possible performance (minAUC column). On the other hand, the ML version of the HSM had poor performance, meaning the data available on a single stereo-pair may be insufficient to train the HSM for good AUC.

It is interesting to note, however, that cost function choice is crucial: note how it had higher impact on the AUC than model choice itself. We argue that the reason for this is that the models presented here are well estimated, rendering their fit to the real distribution, and performance, very similar to each other. Note again in Table 2 and 3 that obtained AUCs are very close to their optimal values, both in the indoors and outdoors set. Since optimal AUC depends on the error rate achieved by each cost function, as shown in [8], then as long as close-to-optimal AUCs are obtained on each model, performance will depend mainly on the cost function. The HSM seems to achieve AUC values that are closer to the optimal for each cost function.

Importantly as well, the results show once more that the usage of the datasets with ground-truth to train parametric models is (not only tedious but also) unnecessary, and our proposed ML strategy for parametric models leads consistently to high performance without the need for GT.

5.4 All models: evaluation on winner-take-all failure

We now present all models' performance regarding $C(d \in GT)_{badpx}$: the confidence given to true disparity when WTA fails. We compare the different models using this criterion in Table 4.

There is a different ranking of models in terms of AUC and C, which suggests that the appropriate choice of model for stereo applications strongly depends on which criterion is to be optimized. However Merrell's model, which had already scored high in the AUC criterion, performed highest in the C criterion using ML estimation (i.e. without the need for training with ground-truth datasets). Such consistency and convenience of ML-estimated Merrell's model makes it a good candidate model for stereo applications.

Regarding the HSM model, its AGT (ground-truth-trained) version performed quite low. Its ML (no-ground-truth) version performed higher, even though it was poor on AUC (Table 3). In the next section we will see how this balance between AUC and C is actually reflected on high performance of both versions of the HSM in practice.

6 Experimental Results on Application to Occupancy Grids

On a second experimental setup we evaluate the different models on a real application, using our occupancy grid method which integrates stereo confidence. In this section we will describe the setup and results, as well as discuss the relation between grid performance and the AUC and C criteria results.

Our grid method assumes static scenes and so the experimental evaluation was also conducted on a dataset with no moving objects: the KITTI residential area dataset "2011_09_26_drive_0079" [26]. The dataset contains 100 synchronized stereo pairs,

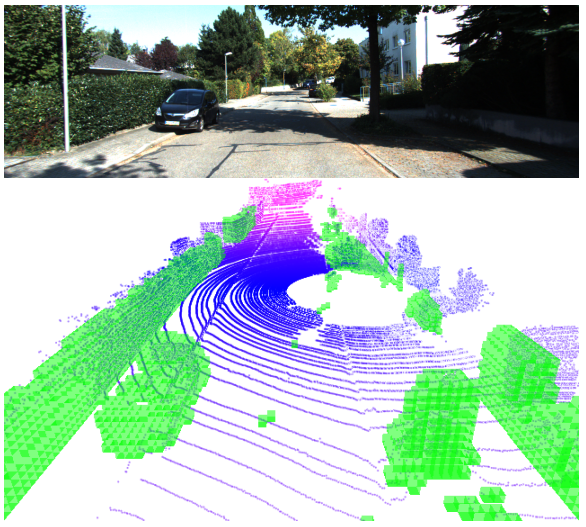

Fig. 6. The KITTI residential area dataset [26] used for occupancy grid evaluation. Green regions on the bottom image represent ground-truth occupied cells. Blue points represent laser data at one of the frames.

laser rangefinder measurements and localization data taken from a moving car, while no moving people or moving cars can be seen. An image of this dataset is shown in Figure 6.

In order to obtain a ground-truth grid, a simple grid algorithm for range data was implemented and run on all frames using the available laser rangefinder data: cells that were occupied with point data in more than a single frame were considered occupied and the rest as free. The localization data, given by the dataset, was assumed to be correct. Cell size used was 20cm x 20cm x 20cm and the resulting grid 60m x 12m x 3m. Generated ground-truth is shown on Figure 6.

To quantitatively evaluate performance of the occupancy grid method we take two measures: "precision" and "recall". Precision measures the fraction of cells classified as occupied which are correct. It is defined as $\frac{tp}{tp+fp}$, where tp (true positives) refers to the number of cells correctly classified as occupied (i.e. occupancy $P > 0.5$) and fp (false positives) refers to the number of cells incorrectly classified as occupied. Recall measures the fraction of occupied cells correctly classified. It is defined as $\frac{tp}{n}$, where n refers to the total number of occupied cells on ground-truth data.

6.1 Model comparison: precision, recall, AUC and confidence on ground-truth

We computed reconstruction performance with all models, including the HSM, using both ABP/AGT and ML parameter estimation. Results are shown in Figure 7. For the ABP parameters of parametric models, we ran the experiment with both maxC and minAUC parameters (see Table 1). Their curves are similar, though, and so we include only one of them (minAUC) in Figure 7. Each dot in the figure represents one instant of time of the image sequence (i.e. frame) and hence an update of the occupancy grid. The first frames are marked with "t=0". Frames used were: 0, 5, 10, etc, in multiples of 5.

The curves in Figure 7 show how the occupancy grid algorithm leads to increasingly higher recall and precision rates as new frames are processed. Precision rates of around 0.9 and recall 0.5 are achieved by most models by the end of the experiment. Another observation is that precision increases slightly with window size, which is consistent with the results in Section 5.

Importantly, the HSM and Merrell models lead to the highest final precision results across most cost function and window size combinations, with the exception of BTSAD 5x5. The ML-estimated exponential had slightly higher precision in that case, however at the cost of low recall. Also note that the HSM model's curve is above other curves during most of the image sequence, showing highest precision, although this distance decreases as the number of used images increases. Models with ML and ABP parameters perform similarly for each model-cost combination, with the exception of Matthies' and the exponential models where ML leads to higher precision but lower recall. These results are consistent with Tables 3 and 4: HSM and Merrell were best performing in either the AUC or C criterion, also ML Exp and Mat had lower C score than their ABP versions, corresponding to the lower recall in the grid application. Overall, higher C criterion is related to higher final grid recall (correlation $r = 0.29$), but not related to precision in our method. Lower AUC is also related to higher final grid recall (correlation $r = -0.35$) and higher final precision (correlation $r = -0.48$).

An interesting observation is how the ML HSM lead mostly to the same performance as the AGT one, even though AUC in the ML case was poor. As we discussed in Section 5.3, the fact that an ML HSM is computed from a single stereo pair could lead to a sparsely populated histogram: thus leading to a poor AUC because the confidence function is not continuous (and ranking of pixels as a function of error rates will also not be continuous). However, the ML histogram is trained from costs at WTA disparity where left-right disparity is consistent. Thus the reason for the ML model's poor AUC could be its bad conditioning near cost values where errors are common (and thus left-right consistency is often not met), even though conditioning is good around common cost values of true disparity. This would explain the still high $C(d \in GT)_{badpx}$ result of the model (see Section 5.4, Table 4), as well as its good performance in the occupancy grid application. Such observations again stress the need for criteria other than AUC for stereo confidence model evaluation, depending on the application.

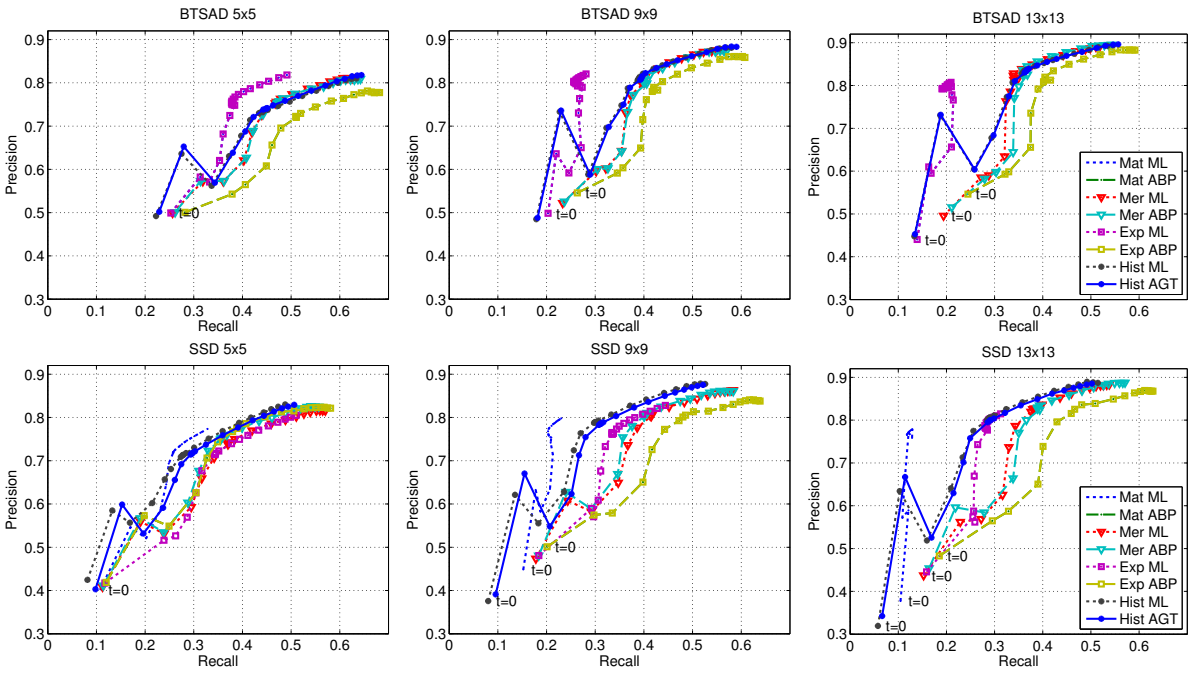

Fig. 7. Comparison of the performance of all models along time when used with the occupancy grid algorithm. Each point represents a different instant of time, while the first frame of the image sequence is marked with "t=0". "Mat ABP" overlaps perfectly with "Exp ABP" on both cost functions, and "Mat ML" overlaps perfectly with "Exp ML" for the BTSAD cost function.

Finally, in Figure 8 we show the reconstruction of ML HSM and Merrell's models (using BTSAD 13x13). The HSM's higher recall can be seen quite clearly (e.g. the car and tree are better reconstructed), although the number of false positives is also slightly higher (since recall is higher and precision rate is not 1).

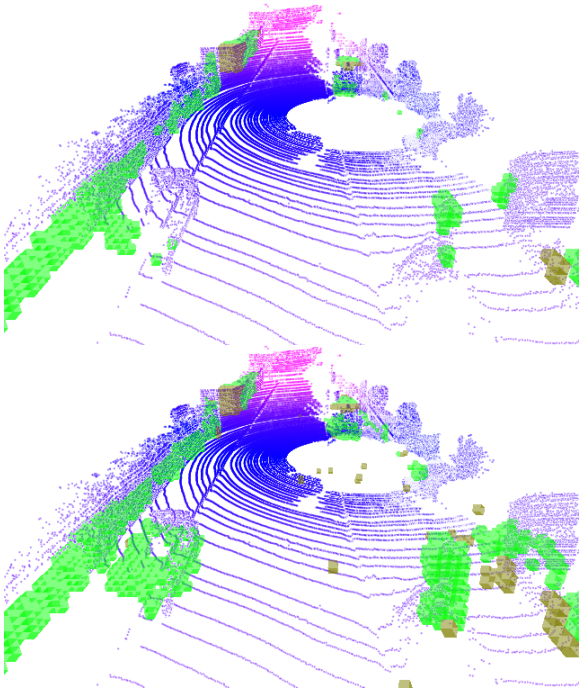

Fig. 8. Reconstruction results obtained using a BTSAD 13x13 cost function with the two top models: Merrell's model (top) and the HSM (bottom). Green squares represent true-positives (i.e. cells correctly classified as occupied), brown squares represent false-positives (i.e. cells incorrectly classified as occupied).

7 Conclusions and Discussion

In this article we evaluated several existing models of confidence which are defined at the whole disparity range. We proposed a new stereo confidence measure, the Histogram Sensor Model (HSM), which consists of a histogram of costs and improves performance in several criteria (i.e. AUC, application to occupancy grids). We also proposed a method to estimate parametric models' parameters that avoids the need for training with ground-truth data. All models were evaluated when applied to two different cost functions (SSD and BTSAD) at different window sizes and model parameters. Contrary to previous stereo confidence measure benchmark literature, we evaluate the models not only using the WTA-relevant criterion AUC, but also with a whole-cost-curve-relevant criterion $C(d \in GT)_{badpx}$: the confidence given to ground-truth on WTA fail. Finally we evaluated the models on a real-world application using a recent global formulation of 3D reconstruction through occupancy grids. Our experimental results lead to several conclusions:

- Performance of parametric confidence measures varies drastically with parameter choice, con-

cretely showing a cliff-maximum-and-tail of performance with parameters. This also leads to the conclusion that over-shooting of parameters is safer than under-shooting. The reason for performance drop when parameters are under-estimated is clear: since the analyzed confidence functions are normalized exponentials of costs, they tend to a min function as the cost normalizer tends to zero (is under-estimated) - leading to a single confidence maximum equal to 1.

- Our results indicate that it is possible in certain applications to train parameters of the parametric models from off-the-shelf datasets with ground-truth disparity (i.e. using average best performing parameters, ABP). However, care should be taken such as to re-train the parameters every time costs, prefilters or dataset conditions are changed.
- We proposed a systematic parameter estimation method for parametric models using maximum likelihood (ML), eliminating the need for any ground-truth or offline training. Our results indicated that these parameters lead to performance in stereo which is similar but slightly closer to the optimum when compared to ABP parameters - which require training datasets with ground-truth. At the same time, the proposed method is trivial to implement and computationally inexpensive. ML should allow for better compensation of environment changes and be more practical when different cost or prefiltering options are applied during the design stage of algorithms.
- The AUC criterion usually compared in the benchmarking literature was shown to be less informative than desirable when used to choose the best model for a global method integrating confidence measures (Cost-Curve Occupancy Grid [6]). We here proposed another criterion, $C(d \in GT)_{badpx}$, which is related to the recall of the grid and ML HSM's performance. Training of parameters by optimizing $C(d \in GT)_{badpx}$ is also subject to lower inter-image variance than AUC.
- In the occupancy grid application the HSM and Merrell's models performed best in terms of grid precision. The HSM actually achieved higher precision earlier on (i.e. using a fewer number of stereo pairs). On the other hand, the exponential and Matthies' models with ABP parameters lead to overall high recall rates but lower precision.
- The HSM was the best performing model in terms of AUC and occupancy grid precision when trained on off-the-shelf datasets with ground-truth. As seen by the shape of the HSM (Figure 3), the distribution of costs at true disparity is not well approximated by a distribution of the exponential-family. We believe this to be a good sign for a push in stereo research towards non-parametric confidence models.
- For applications where AUC is an important criterion, our results show however that the HSM should not be trained on WTA disparity with few data. Merrell's model with ML parameters is a good choice when ground-truth datasets are not available for training, since it scores high in terms of AUC, $C(d \in GT)_{badpx}$ and grid performance.

Important directions of research include new non-parametric models of stereo confidence, or models with low parameter sensitivity. We hope to have made clear that more research into methods for online (no ground-truth) estimation of model parameters has the potential for high impact on stereo and its applications. Other approaches to training the HSM without ground-truth may also be worth investigating, as is the combination of different confidence measures [34].

ACKNOWLEDGMENTS

We deeply thank the reviewers of this article for their invaluable comments and suggestions for improvement of the research. This study was conducted as part of the Research Institute for Science and Engineering, Waseda University, and Humanoid Robotics Institute, Waseda University. It was also supported in part by JSPS KAKENHI (Grant Number: 24360099 and 25220005), Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation, JSPS, Japan, and the EU Project Poeticon++ FP7-ICT-288382.

REFERENCES

[1] J. Sun, N. Zheng, and H. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 1–14, 2003.

[2] D. Scharstein and R. Szeliski, "Stereo Matching with Nonlinear Diffusion," *International Journal of Computer Vision*, vol. 28, no. 2, pp. 155–174, 1998.

[3] C. J. Pal, J. J. Weinman, L. C. Tran, and D. Scharstein, "On Learning Conditional Random Fields for Stereo," *International Journal of Computer Vision*, vol. 99, no. 3, pp. 319–337, Oct. 2010.

[4] R. A. Newcombe, S. Lovegrove, and A. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2320–2327.

[5] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[6] M. Brandao, R. Ferreira, K. Hashimoto, J. Santos-Victor, and A. Takanishi, "On the formulation, performance and design choices of cost-curve occupancy grids for stereo-vision based 3d reconstruction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2014.

[7] J. Cech and R. Sara, "Efficient sampling of disparity space for fast and accurate matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[8] X. Hu and P. Mordohai, "A Quantitative Evaluation of Confidence Measures for Stereo Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[9] G. Egnal, M. Mintz, and R. P. Wildes, "A stereo confidence metric using single view imagery with comparison to five alternative approaches," *Image and vision computing*, vol. 22, no. 12, pp. 943–957, 2004.

[10] H. Hirschmüller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1582–99, Sep. 2009.

[11] L. Matthies, T. Kanade, and R. Szeliski, "Kalman filter-based algorithms for estimating depth from image sequences," *International Journal of Computer Vision*, vol. 236, pp. 209–236, 1989.
[12] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 9, pp. 920–932, 1994.
[13] a. Fusiello, V. Roberto, and E. Trucco, "Efficient stereo with multiple windowing," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, no. 2, pp. 858–863, 1997.
[14] L. Matthies and M. Okutomi, "A Bayesian foundation for active stereo vision," *Proc. SPIE Sensor Fusion II: Human and Machine Strategies*, pp. 1–13, 1989.
[15] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 401–406, 1998.
[16] P. Mordohai, "The self-aware matching measure for stereo," in *IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 1841–1848.
[17] R. Mayoral, G. Lera, and M. J. Perez-Ilzarbe, "Evaluation of correspondence errors for stereo," *Image and Vision Computing*, vol. 24, no. 12, pp. 1288 – 1300, 2006.
[18] A. Torabi, M. Najafianrazavi, and G. A. Bilodeau, "A comparative evaluation of multimodal dense stereo correspondence measures," in *2011 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, 2011, pp. 143–148.
[19] M. Gong and Y.-H. Yang, "Fast unambiguous stereo matching using reliability-based dynamic programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 998–1003, 2005.
[20] C. Dima and S. Lacroix, "Using multiple disparity hypotheses for improved indoor stereo," in *IEEE International Conference on Robotics and Automation*. IEEE, 2002, pp. 3347–3353.
[21] N. Sabater, A. Almansa, and J. M. Morel, "Meaningful matches in stereovision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 930–942, May 2012.
[22] R. Sára, "Finding the largest unambiguous component of stereo matching," in *Proceedings of the 7th European Conference on Computer Vision-Part III*, ser. ECCV '02. London, UK, UK: Springer-Verlag, 2002, pp. 900–914.
[23] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the power of stereo confidences," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 297–304.
[24] M. Brandao, R. Ferreira, K. Hashimoto, J. Santos-Victor, and A. Takanishi, "Integrating the whole cost-curve of stereo into occupancy grids," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, November 2013, pp. 4681–4686.
[25] R. Shade and P. Newman, "Choosing where to go: Complete 3D exploration with stereo," *2011 IEEE International Conference on Robotics and Automation*, pp. 2806–2811, May 2011.
[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.
[27] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.
[28] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian Conference on Computer Vision*, 2010.
[29] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.
[30] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.
[31] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. I–195–I–202, 2003.
[32] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.
[33] H. Hirschmuller and D. Scharstein, "Evaluation of Cost Functions for Stereo Matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Ed., 2007, pp. 1–8.
[34] R. Haeusler and D. Kondermann, "Synthesizing real world stereo challenges," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, J. Weickert, M. Hein, and B. Schiele, Eds. Springer Berlin Heidelberg, 2013, vol. 8142, pp. 164–173.

Martim Brandão received his M.Sc. in Electrical and Computer Engineering from Instituto Superior Técnico (IST, Portugal) in 2010. He was Research Assistant at the Computer and Robot Vision Lab (IST) in 2011, and Research Student at Takanishi Laboratory (Waseda University, Japan) until 2013. He is now a Ph.D. student at Waseda University. His research focuses on computer and robot vision topics such as 3D reconstruction, visual tracking and robot motion planning.

Ricardo Ferreira received his B.Sc. in Electrical and Computer Engineering in 2004, M.Sc. in 2006 and Ph.D. in 2010, all at Instituto Superior Técnico (IST). In his M.Sc. he studied underwater stereo reconstructions of 3D scenes when observed through an air-water interface and the Ph.D. was focused on reconstructing paper-like surfaces from multiple camera images. His research interests include manifold optimization and geometric problems in robotics and computer vision.

Kenji Hashimoto is an Assistant Professor of the Research Institute for Science and Engineering, Waseda University, Japan. He received the B.E. and M.E. in Mechanical Engineering in 2004 and 2006, respectively, and the Ph.D. Integrative Bioscience and Biomedical Engineering in 2009, all from Waseda University, Japan. His research interests include walking systems, biped robots, and humanoid robots.

Atsuo Takanishi is a Professor of the Department of Modern Mechanical Engineering, Waseda University and director of the Humanoid Robotics Institute (HRI), Waseda University, Japan. He received the B.S.E. in 1980, M.S.E. in 1982 and Ph.D. in 1988, all in Mechanical Engineering from Waseda University. His current research is related to Humanoid Robots and its applications in medicine and well-being.

José Santos-Victor received a Ph.D. in Electrical and Computer Engineering in 1995 from Instituto Superior Técnico (IST, Portugal), in Computer Vision and Robotics. He is Associate Professor at the Department of Electrical and Computer Engineering of IST and a researcher of the Computer and Robot Vision Lab. His is interested in Computer and Robot Vision, particularly visual perception and the control of action, and biologically inspired vision and robotics.